**비즈니스 모델링 2**

**Assignment 2**

**Due: 11/20, 2020 (금요일), 11:59PM**

**(Total Points: 22)**

**Instructions:** 다음 문제들을 잘 읽고, 요구되는 코드를 작성하고 설명하시오. 코드와 설명을 포함한 모든 답안을 ‘HW2\_김일주(자기이름).r’ 하나의 파일에 저장하여 제출하시오. 설명을 요하는 문제의 경우 주석을 사용하도록 하고, 코드와 설명이 몇번 문제에 대한 답인지를 반드시 표시하시오. 늦은 제출은 받지 않으니, 제출시한에 맞추도록 유의하시오. 반드시 자신들의 이해를 충분히 보이는 방법으로, 자신들의 Words를 사용하여 답안을 작성하시오. 질문있을 시 언제든지 [ijkim@mju.ac.kr](mailto:ijkim@mju.ac.kr)로 연락하시오.

1. **수업시간에 활용한 아침식사용 시리얼 예제의 데이터를 사용하여 아래와 같이 데이터를 탐색하고 요약하시오.**
2. **(3pts) 어떤 변수들이 양적/수치적 변수인가? 순서형 변수는 어떤 것인가? 명목형 변수는 어떤 것인가? 데이터를 보고 설명하시오.**

**양적/수치적 변수는 calories, proteinm fat, sodium, fiber, carbo, sugars, potass, vitamins, weight, cups 까지의 변수이며, 순서형 변수는 rating변수이다.**

**명목형 변수는 name, mfr, type, shelf이다.**

1. **(3pts) 각각의 양적 변수에 대해서 평균, 중앙값, 최솟값, 최댓값 그리고 표준편차를 계산하시오. 이것은 R의 sapply() 함수를 통해 계산될 수 있음 [예를 들어, sapply(data, mean, na.rm=TRUE)].**
2. **(3pts) R을 사용하여 양적 변수 각각에 대한 히스토그램을 작성하시오. 히스토그램과 요약 통계량을 바탕으로 다음 질문에 답하시오.**
3. **어떤 변수의 변동이 가장 큰가?**

**Soduim이 0~320까지로 변수들 사이에서 값의 변동이 가장 크다.**

1. **어떤 변수가 치우쳐 있는가?**

**Fiber 가 0~4까지의 값이 69개로 가장 많이 치우져 있다.**

1. **극단값으로 보이는 값이 있는가?**

**Fiber의 0~2의 개수가 다른 그래프들과 비교했을 때 최소 48개인 극단값으로 보인다.**

1. **(3pts) R을 이용하여 저온용과 고온용 시리얼에 함유된 칼로리를 비교하기 위해 박스플롯을 병렬로 나란히 그리시오. 이를 통해 무엇을 알 수 있었는지 쓰시오.**

**고온용 시리얼의 칼로리는 100으로 일정했는데 저온용 시리얼은 칼로리의 범위가 넓은 것을 알 수 있었다. 제일 낮은 칼로리 50부터 해서 가장 높은 칼로리는 160까지 나왔다. 하지만 대부분은 100 ~110에 분포하고, 평균은 107 이라는 것을 알 수 있다.**

1. **(3pts) R을 사용하여 시리얼이 전시된 진열대 높이에 따라 소비자 평점의 박스플롯을 병렬로 나란히 그리시오. 이를 통해 만약 진열대 높이로부터 소비자 평점을 예측할 수 있는 것으로 보인다면, 진열대 높이의 세 가지 범주를 그대로 유지할 필요가 있겠는가에 대해 이야기해 보시오. 소비자 평점을 예측하기 어렵다면 왜 그런지 설명하시오.**

**Shelf == 1일때의 평균값은 46.1이고 shelf == 2일떄의 평균값은 34.9, shelf == 3일 때의 평균값은 45.2가 나왔다. 만약에 shelf의 높이에 따라 소비자의 평점을 예측한다면 shelf ==1과 shelf == 3일 떄의 소비자 평점이 비슷하기 때문에 둘을 굳이 나눌 필요는 없어보인다. 하지만 선반의 높이와 소비자 평점 사이에서 비례 혹은 반비례적으로 상관관계가 생기지 않기 때문에 소비자 평점을 예측하기에는 어려워 보인다.**

1. **(3pts) 양적 변수에 대한 상관계수 표를 작성하시오 [cor() 함수 사용]. 또한 이들 변수에 대한 산점도 행렬을 작성하시오 [plot(data) 함수 사용]. 그리고 다음 질문에 답하시오.**
2. **어느 변수 쌍이 가장 높은 상관관계를 나타내는가?**

**Fiber-potass이 0.91로 가장 높은 상관관계를 가진다.**

1. **이들 상관관계를 바탕으로 어떻게 변수의 개수를 축소할 수 있는가?**

**Fiber와 potass가 서로 강한 양의 상관 관계를 가지므로 둘 중 하나의 변수를 누락(제거)하는 것을 고려할 수 있다.**

1. **데이터를 먼저 정규화하였다면 상관관계가 어떻게 변하겠는가? (수업시간에 이야기를 못하고 넘어갔으니, 점수와 상관없이 대답해 보시오.)**

**데이터를 정규화하면 NA값이 사라지기 때문에 상관관계가 높아질 것 같다.**

1. **(4pts) Ch4. 강의슬라이드 22페이지의 Table 4.11에서 13개의 수치형 변수에 대한 첫 번째 주성분(principal component)에 대하여, 이 주성분무엇을 나타내는지, 그리고 이 주성분의 여러 정보중 왜곡된 것이 있다면 어떤 것인지를 수업내용에 기반하여 설명해보시오**

**Table 4.11은 시리얼에 들어있는 변수들의 수치를 PCA한 것이다.**

**첫번째 주성분은 몇 %의 정보를 차지하는지를 보여준다. 첫번째의 정보만 포함되면 54%의 정보가 포함된다.**

**왜곡된 정보는 Sodium와 patass가 왜곡된 정보이다. 그 이유는 나머지 변수들은 gram기준인데 sodium은 milligram이기 때문이다. 그래서 98%정도의 비율을 차지한다고 나와있지만 실제로는 틀린 값이다. 그래서 Normalized를 하면 sodium의 첫번째 주성분이 18%정도가 나오고, patass의 두번째 주성분은 30%정도가 나온다.**